BIOPRESERVATION AND BIOBANKING
Volume 14, Number 5, 2016
Mary Ann Liebert, Inc.
DOI: 10.1089/bio.2015.0100

Reconsidering Anonymization-Related Concepts and the Term "Identification" Against the Backdrop of the European Legal Framework

Murat Sariyar[1,2] and Irene Schlünder[2]

Sharing data in biomedical contexts has become increasingly relevant, but privacy concerns set constraints for free sharing of individual-level data. Data protection law protects only data relating to an identifiable individual, whereas ''anonymous'' data are free to be used by everybody. Usage of many terms related to anonymization is often not consistent among different domains such as statistics and law. The crucial term ''identification'' seems especially hard to define, since its definition presupposes the existence of identifying characteristics, leading to some circularity. In this article, we present a discussion of important terms based on a legal perspective that it is outlined before we present issues related to the usage of terms such as unique ''identifiers,'' ''quasi-identifiers,'' and ''sensitive attributes.'' Based on these terms, we have tried to circumvent a circular definition for the term ''identification'' by making two decisions: first, deciding which (natural) identifier should stand for the individual; second, deciding how to recognize the individual. In addition, we provide an overview of anonymization techniques/methods for preventing re-identification. The discussion of basic notions related to anonymization shows that there is some work to be done in order to achieve a mutual understanding between legal and technical experts concerning some of these notions. Using a dialectical definition process in order to merge technical and legal perspectives on terms seems important for enhancing mutual understanding.

Keywords: anonymization, data protection, identity, re-identification

Introduction

Making research data available to the scientific community for future research purposes is often postulated and required by funding policies, where public funds are involved.[1–3] Sharing data has many benefits, for example, ensuring the validation of results, receiving feedback to improve data quality for ongoing data collection efforts, and facilitating innovative secondary analyses and meta-analyses on the original data.[4,5] Even though the research community widely acknowledges the necessity and usefulness of making data collections available for use within new and different contexts, there are two main issues regarding such secondary use. First, there is a legitimate interest of researchers collecting data and harvesting it to such a degree that their efforts seem worthwhile. Second, privacy concerns raise legal issues, since the data of patients and donors are at stake, and these data are very often sensitive data. We will deal here with this second key barrier for sharing research data in Europe.

The secondary use of data relating to individuals is allowed under most legislations if corresponding consent is available or the national law provides for special permissions to use the data or the data are anonymized.[6] Anonymization makes data sharable without further data protection constraints such as consent management, which produces an additional, sometimes significant, administrative burden, and which individual researchers or research groups may not be able or have the resources to take on. This is one of the main reasons why different scientific communities (e.g., the database community, the statistical disclosure community, and the cryptography community) are developing techniques and methods for anonymization.[7–13]

However, anonymization can have a negative impact on the use of the data for the research community. Distortions caused by anonymization techniques should therefore not be accepted as a prerequisite to make data sharable. On the contrary, application of anonymization must be justified by well-established rules that serve a public interest. The

[1]Institute of Pathology, Charité–University Medicine Berlin, Berlin, Germany.
[2]TMF (Technologie- und Methodenplattform e.V.), Berlin, Germany.

privacy of patients and donors whose data are processed certainly is such a public interest as well as a fundamental right of the respective individuals (e.g., Article 8 European Charter of Fundamental Rights). Data protection law gives substance to this fundamental right and at the same time provides rules for the trade-off with research interests.

We deal here with notions related to identity disclosure risks and especially with the question of what identification could mean concretely from a legal perspective. Some anonymization techniques/methods for preventing reidentification are provided as well.

Background: Legal Framework and Reasons for Anonymization

Anonymization within the legal data protection framework in the EU

Legal constraints for data sharing result from the legally protected interest of information privacy, that is, the ''right to select what personal information about me is known to what people.''[14] Thus, data protection law protects only data relating to an identifiable individual, whereas ''anonymous'' data are free to be used by everybody. This general principle is recognized worldwide, nevertheless data protection law details differ from state to state and the same is true for the legal concepts of anonymity. Thus, it seems to be useful to refer to a concrete legal data protection framework to pave the way in the jungle of terms, notions, and concepts.

According to Article 2(a) of the basic legal framework in Europe, the EU Data Protection Directive of 1995,[15] personal data are ''any information related to a natural person, who is identified or identifiable directly or indirectly in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.'' Recital 26 makes clear that ''the principles of protection shall not apply to data rendered anonymous.'' Thus, anonymization is a means to make data available without further legal constraints, at least with respect to data protection law.

Although there is no definition of *anonymity* in the Directive, the term can be deduced from the term ''personal data,'' being the opposite term to ''anonymous data.'' According to Recital 26: ''to determine, whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person.'' This provision takes account of the fact that absolute anonymity is not achievable, that is, that there will hardly ever be a zero risk of reidentification.[16] Therefore, it is widely accepted that *de facto* anonymity is sufficient and this approach is also pursued by the Article 29 Working Party—being the European body for giving advice on the interpretation of the data protection directive—as many statements indicate.[17]

The term *de facto* anonymity also implies that anonymity is not static: the same information (data set or data record) can be anonymous in one context and personal data in another. The full name (e.g., ''Harry Smith'') might not be an identifier without additional information in an epidemiological database, whereas it is sufficient to be identified in a classroom. In technical terms, this is described as the fact that the full name is not structurally unique, but might only be empirically unique.[18] In the latter case, reference information available to those who have access to the data plays a decisive role in the decision on whether data are identifiable. Therefore, a deidentified data set available to anybody on the Internet should be more protected than the same data set stored on a university server accessible only to a few researchers. This leads to the question whether access policies as additional safeguards can have an impact on the status of anonymity.

Second, whether means are likely reasonably to be used also depend on the motives of an attacker to reidentify a person from a deidentified data set. Data relating to people of public interest require a higher level of protection, and sensitive data require an even higher level. Therefore, the level of deidentification has to be defined accordingly.

Last but not least, the invasion of privacy, which has to be taken into account while assessing the risk of reidentification, depends on the type of information disclosed. For example, the reidentification of a spectacle wearer in a data set in which spectacles purchased from an optician are listed is a disclosure of personal data, for which the invasion of privacy would be quite low. In contrast, the reidentification of an HIV patient could have much more impact on the person's life, be it his family life, working environment, etc.

The concept of anonymization will not change under the General Data Protection Regulation, which is going to replace the current Directive. The definition of personal data given in Article 4 para 1 of the latest version (see result of the Trilogue: www.statewatch.org/news/2015/dec/eu-council-dp-reg-draft-final-compromise-15039-15.pdf) does not substantially differ from that in the Directive. In addition, there is no indication that the concept of anonymization or its impact on the possibility of data sharing has been changed.

Anonymous data under EU law must be distinguished from pseudonymized data, also known as coded-anonymized. In the latter case, artificial identifiers (pseudonyms) replace the most identifying fields. Although there are many advantages in using pseudonymized data instead of anonymized data, even from the perspective of patients and donors, such as the possibility to get feedback regarding research results, pseudonymized data remain legally protected personal data, at least for all who can retrace the individual, for example, for follow-ups. This article only deals with full (*de facto*) anonymity, which does not allow the provision of feedback or to have follow-ups. It is an unresolved question whether pseudonymized data can be considered anonymous for those who have no access to the key (see the decision of the German Federal Court of Justice, which has presented this question to the European Court of Justice: http://dejure.org/dienste/vernetzung/rechtsprechung?Gericht=BGH&Datum=31.12.2222&Aktenzeichen=VI%20ZR%20135/13).

The term ''anonymization'' is not identical to ''deidentification.'' Deidentification is the removal of attributes known to increase the risk of identification, and this can be seen as a preliminary step for producing anonymous data.[19,20] It requires, however, a further assessment as to whether the deidentification process achieves anonymization. Equally, the well-known HIPAA list of identifiers in health data, which have to be removed under U.S. law before sharing the data according to the HIPAA Act, leads only to deidentification, not to anonymous data in the sense of data protection law in Europe.

Balancing privacy and usefulness of the data

The first main goal of anonymization methods is the transformation of data to be released in such a way that it

does not exhibit information of individuals that was not previously known (data protection)[10,13]; the second one is the preservation of as much of the underlying information as possible (usefulness).

The technical approach is to balance the two goals by defining the desired end result[21]: a *minimal anonymous* data set satisfies a given privacy requirement by applying a set of anonymization operations that cannot be reduced without violating the requirement (just focusing on the number of steps, irrespective of how useful the data might be according to some utility metric). In contrast, an *optimal anonymous* data set satisfies the given privacy requirement and contains the highest amount of information among all privacy-satisfying data sets according to a chosen utility metric (e.g., instead of perturbing one important variable, two rather unimportant variables are perturbed). Utility metrics are necessary for assessing the usefulness of the data. For undefined data usages, general-purpose metrics such as the information entropy[22] and coverage of the original domain attributes[23] can be used. If the use cases are specified concretely, special-purpose metrics can be used, for example, for assessing the usefulness of the data for a classification task with methods such as logistic regression, association rules, or classification trees. Here, results on the original and the modified data are compared (errors made, variance, etc.).

Although it seems rather easy to define the general goal of anonymization from a technical perspective, it is far more difficult to decide a given case from a legal perspective, for example, to what extent has an electronic health record of patients to be deidentified to share the data within a certain research project? Would the situation change if the data are included in a database for future research purposes? Can genetic data be kept in? How is the situation to be assessed if the database integrates patient data from other sources? Whereas anonymization methods often start from the assumption of a static result, that is, after the anonymization process is anonymous or not, the legal view is different: a risk assessment has to be made for every single situation. The trade-off between data protection on the one hand and usability of the data for the respective research purpose on the other hand has to be made again as soon as the situation has changed. Those changes can have very different reasons: reidentification techniques have been enhanced over time, the data have been enriched by integrating other sources, some of the patients involved have become of interest for the public or potential attackers, reference information has increased, etc. The latter can, for example, be observed regarding genetic information and biosamples.

The concrete assessment of the usefulness of data from a legal perspective is frequently only possible through guidance by a group of experts. One has to rely on statisticians or other data analysts in their judgment regarding the usefulness of data.

Disclosure risks

In the method-oriented literature, three main types of identity disclosure risks are discussed: identity, attribute, and membership disclosure.[24] *Identity disclosure* is usually regarded as the singling out of an individual within a data set, which means that all information contained in the data set about this individual is revealed. *Attribute disclosure* is the unveiling of sensitive information of an individual, e.g., having a specific disease, without performing a singling out. For example, the individual can be linked to a set of rows with a common value for the sensitive attribute (see the short discussion on Table 2). Finally, *membership disclosure* means that an attacker is able to determine whether an individual is contained in a data set. In contrast to attribute disclosure, the attacker only knows that a specific individual is in the data set without the ability to deduce the concrete values of sensitive attributes. All these disclosure types necessitate slightly different forms for measuring risks. For example, for assessing the reidentification risk, the *population uniqueness* is frequently used, which can be estimated by the proportion of records that are unique in the original data set/population.[25]

Irrespective of the distinctions already given, assessing disclosure risks requires consideration of several aspects, especially the following:

Properties of the data: structure of the data (relational data: one row per individual or transactional data, for example, follow-up time-to-event data), the data type (e.g., strings and numerical values), scope of the data (e.g., clinical, survey, and genomic data), number of attributes, and so on.

Type of user: for example, researcher, nonexperts, or machines.

Type of application: for example, not fixed or statistical analyses like regression, classification, or clustering.

Type of access: open access or some type of restriction.

Modus of release: for example, release of a subset (horizontal partition) or release of different kinds of characteristics (vertical partition).

Attacker model: at least two kinds of attackers should be discerned to assess the related disclosure risk. The first one is the ''prosecutor,'' who targets one specific individual, and the second is the ''journalist,'' who wants to discredit the institution that is issuing the data and targets any individual.

Basic Terms: Identifier, Quasi-Identifier, and Sensitive Attribute

As a background for the terms examined here, a simplified example is given. A raw database of a hospital (Table 1) consists of IDs, sex, year of birth, ZIP code, and the ICD-10 code. The hospital wants to release the database to enable some statistical analyses for the general audience. Before publishing, the data are transformed in such a manner that at least two records have identical values for the so-called quasi-identifiers (QIDs) (Table 2), and this kind of anonymization is called k-anonymization.

The European data protection law does not provide a classification of attributes related to anonymization, but only refers to the whole set of attributes as either identifying (personal data) or not (anonymized data). It protects ''sensitive'' data by imposing higher constraints, but only so far, as it is personal data.

Unique identifiers

An identifier is a unique identifier only if a single specified individual is associated with it. Examples are social

Table 1. Original Data Set Without Any Key Identifiers

	QIDs			*SensAttr*
Irrelevant ID	*Sex*	*Year of birth*	*Zip code*	*ICD-10 code*
6	M	1980	10117	Q90.1
8	F	1966	10117	F31.1
1	M	1979	10118	F31.0
9	M	1988	11067	F31.9
11	F	1965	11910	G50.1
4	F	1983	11934	F34.8
10	M	1973	12002	F34.8
3	F	1967	12033	F31.9
2	M	1989	12200	F31.1
5	F	1959	12200	G50.1
12	M	1976	13011	Q90.1
7	M	1975	13135	Q90.0

ID is an irrelevant number for record identification. QIDs are sex, year of birth, and zip code. The sensitive attribute is the ICD-10 code.

F, female; M, male; SensAttr, sensitive attribute; QIDs, quasi-identifiers.

security numbers and biometric identifiers, including fingerprint, retinal and voiceprints, some genetic data, etc. The ID in Table 1 is not a globally unique identifier, but only an artificial identifier for the data set. In real-life settings, such an identifier should be masked as well, because several releases of an anonymized data set with an identical ID increase the probability of reidentification (such a stable ID across several data sets makes the data set a pseudonymized data set, irrespective of other safeguards).

In contrast, biometric and genetic identifiers are interesting for statistical analysis and can be highly sensitive. In addition, they are not changeable or erasable as artificial identifiers. Once attributed to the respective individual, they will serve to disclose his or her identity forever. The strongest biometric identifier seems to be the passport photo, since it is ''readable'' for everyone and thus very easily attributable, whereas a full genome alone is unlikely to be meaningful without the support of technical devices.

Table 2. Transformed Data Set

	QIDs			*SensAttr*
Irrelevant ID	*Sex*	*2-decade-range of birth*	*Zip code*	*ICD-10 code*
6	M	1970–1989	1011*	Q90.1
8	F	1950–1969	10117	F31.1
1	M	1970–1989	1011*	F31.0
9	M	1970–1989	11067	F31.9
11	F	1950–1969	1191*	G50.1
4	F	1970–1989	1193*	F34.8
10	M	1970–1989	12*	F34.8
3	F	1950–1969	12*	F31.9
2	M	1970–1989	12*	F31.1
5	F	1950–1969	12*	G50.1
12	M	1970–1989	13*	Q90.1
7	M	1970–1989	13*	Q90.0

Records 8 and 9 are suppressed because they would lead to a coarsening of four other records in their zip code. For ''year of birth,'' a full-domain generalization to ''2-decade-range of birth'' is made, whereas local generalizations (or recodings) for the zip code are made to achieve a two-anonymous data set with minimal distortion.

*Indicates that digits were omitted.

Quasi-identifiers

QIDs are ''variable values or combinations of variable values within a dataset that are not structurally unique but might be empirically unique and therefore in principle uniquely identify a population unit.''[18] QIDs should contain attribute A if an attacker could potentially obtain A from other external resources, if it is used for data analysis, and if it has discriminatory value. Examples of QIDs are gender, age, postal codes, race, ethnicity, etc.

QIDs are at the heart of all efforts to protect privacy. They are the attributes that are mainly addressed by the phrase ''to determine, whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person.'' Background knowledge and external data sources related to the QIDs can increase the probability of reidentification drastically. Names are also QIDs, even though they are often called ''direct'' identifiers, because they are not structurally unique; however, like unique identifiers they are in most cases uninteresting for analyses and can be deleted in the data set.

Sensitive attributes

According to Article 8 of the Data Protection Directive, special categories of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and data concerning health or sex life are subject to a higher level of protection. These categories are described as ''sensitive data.'' The special protection of these categories of personal data reflects the fundamental rights to nondiscrimination as per Article 14 of the European Convention on Human Rights. It is a basic principle of a democratic and open society that people should exclusively be judged and treated according to their free behavior—and not by inherited or otherwise unchangeable attributes or by their religious or political convictions.

So, a sensitive attribute is a variable that contains personal information with a high impact on privacy, and for which it is required that no attacker has knowledge from other sources than from the data subject themselves. Machanavajjhala et al.[26] state that ''an attribute is marked sensitive if an adversary must not be allowed to discover the value of that attribute for any individual in the dataset.'' Hence, the technical perspective seems to coincide with the definition of the legal perspective. However, the definition of sensitive attributes in the research community does frequently not often refer to any specific law,[10] but classifies further attributes as sensitive in specific applications, whereby it remains unclear on which basis this classification is done (e.g., is salary a sensitive attribute?). The consequence of defining an attribute as sensitive is that in addition to the protection of identity disclosure by means of sufficiently masking QIDs, measures have to be taken for protecting against attribute disclosure (e.g., in addition to k-anonymity, l-diversity has to be achieved).

Often, an implicit distinction between QIDs and sensitive attributes is made that relates to the assumption of whether external information on them is possible or not.[21] If it is assumed that no external information (i.e., information outside the context of gathering/producing the data) is available, then that attribute cannot be used for increasing the disclosure probability, and hence can be published without the necessity of distortion (if everything else is anonymized appropriately). Because sensitive attributes are frequently the main endpoints for statistical analyses, such a nonperturbed sharing of sensitive attributes is highly desirable. However, whether an attacker can have external knowledge about an attribute cannot be determined in advance, especially when parameters of the context are changing. In most cases, therefore, sensitive attributes should be considered as QIDs with special concerns regarding attribute disclosure risks, which, however, can have severe detriment effects on the usefulness of the data, when anonymization methods are applied.

Even if an attacker cannot point to a record as belonging to one person, there could be the possibility to infer sensitive information about one person. In Table 2, records with ID 12 and 7 are indistinguishable with respect to the QIDs and are different with respect to the ICD-10 code, which represents the sensitive attribute in this case. However, an attacker can still infer that someone who was treated in the hospital, was born in 1986, is male, and lives in the 13000 zip code area must have the Down syndrome.

What Is Identification?

While discussing disclosure risks in the preceding section, the term ''identification'' has been used without proper definition. It seems obvious what this term means, but closer scrutiny reveals that the precise meaning of identification of a person is far from being self-evident. The Article 29 Data Protection Working Party states: ''Identification is normally achieved through particular pieces of information which we may call 'identifiers' and which hold a particularly privileged and close relationship with the particular individual. Examples are outward signs of the appearance of this person, like height, hair color, clothing, etc., or a quality of the person which cannot be immediately perceived, like a profession, a function, a name, etc.''[27] The Directive mentions those identifiers in the definition of personal data in Article 2 when it states that a natural person ''can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.''

Does such a definition really help in understanding what identification is? What does it mean that an identifier has to be associated with an individual and data about him? In philosophical debates, the ship of Theseus is an archetypical example of the questions: how many and what kinds of changes are possible before losing the identity of something? Before I can associate an identifier to an individual, I have to identify him, and this means: I have to determine under which circumstances I would associate the same identifier to an individual. Pondering about this problem, the circularity of the definition already described becomes manifest. I can only associate an identifier to an individual if I already have identified him or her. To avoid circularity, one solution consists in equating one unique (and relative stable) identifier with the individual, for example, the (genomic) fingerprint. Identification would then mean to associate other identifiers (the question remains: what kind of identifiers?) with the fingerprint. For some unique identifiers, for example, the combination of full name, date of birth, place of birth, and passport photo, such an association seems unnecessary, but it is useful, for example, because associating these attributes with address details or fingerprints allows finding nonpresent individuals who had undergone some changes in their appearances.

Another question is related to the contexts of identification. According to the Article 29 Data Protection Working Party opinion on personal data, ''a natural person can be considered as 'identified' when, within a group of persons, he or she is 'distinguished' from all other members of the group.''[27] Hence, it is relevant to what kind of groups one refers. When the group is limited or predefined in the sense that not the whole world population is included, one can speak of ''relative identification.'' In this case, it is enough that the person is singled out (''distinguished'') within the group. If the group is the whole world population, one can speak of ''absolute identification.'' Such a global singling out allows, for example, traceability, if the distinguishing features are found elsewhere again. Our focus here is on the second type of singling out.

We refrain from equating singling out with identification. Of course, singling out is the basic requirement to identify an individual, but is this sufficient? In that case, every unique identifier would be enough to identify a person. Knowing the full genome of an individual would therefore be sufficient to identify him or her. However, this seems to be at odds with the usual understanding as well as with the legal perspective. Normally, no one assumes that the knowledge of the full genome or a fingerprint as such reveals the identity of an individual. To know somebody in ''flesh and bones''[27] seems to require more than the knowledge of a unique identifier. As already stated (a solution to prevent circular definition of ''identification''), it is the associations of a unique identifier with other personal information that should be considered as ''identification.''

Information that serves to identify a person in everyday life situations is contained in a personal ID card: the full name, date of birth, place of birth, height, nationality, and passport photo. None of these particular attributes can be qualified as a unique identifier, even though a passport photo is a strong identifier, which is nevertheless not wholly stable. These attributes must be available and understandable to the ''average individuals'' (no attackers). The question is: would someone who already knows the person by having met him in ''flesh and bones'' be able to attribute the respective information to him? When associating such information with artificial unique identifiers, circularity of identification seems to be unavoidable (how to be sure that one is only pretending to be the one to whom the identifier belongs?). Therefore, it is crucial to have a natural unique identifier, to which such personal information is associated.

In conclusion, identification of an individual (based on digital data) means to know a globally unique natural

identifier (allowing a singling out), which can be a combination of attributes and to associate them with a set of attributes. Both together allow a singling out and the recognition of the individual. The genome and fingerprints are unique identifiers, and they can stand for the individual, but they are not sufficient to identify an individual without further reference information. In general, there is no fixed set of (reference) information that is always sufficient for identification.

How Can Reidentification Be Prevented?

Anonymization methods for mitigating disclosure risks can be classified into several different ways.[10,12,13] In contrast to many unstructured listings and bipartitions such as by LeFevre et al.,[28] we propose the following tripartition:

(1) methods for hiding attribute details (e.g., generalization and suppression),
(2) methods that disassociate attributes (e.g., anatomization and disassociation), and
(3) methods that perturb the data (e.g., randomization and permutation).

A review on anonymization methods that list them according to the different disclosure risks is given by Gkoulalas-Divanis et al.[29] We opt here for a classification that is based on the techniques underlying the methods, because we want to give an orientation regarding the central ingredients for the methods.

Hiding attribute details: generalization and suppression

Fulfilling the k-anonymity criteria, which focuses on reducing the reidentification risk, is the most targeted goal within this group of methods.[9] k-Anonymity requires that all equivalence classes (an equivalence class is the list of all records that have the same values for the QIDs) have at least a size of k. In this case, the reidentification probability has an upper bound of 1/k. When using generalization for categorical variables, usually a domain hierarchy taxonomy must be generated (for alternatives, see Ref.[30]). The most granular level is at the bottom, and every higher hierarchy level coarsens the data, for example, 45-year-olds (at the bottom) → [50, 60] interval → [40, 70] interval, etc. Generalization is then the replacement of attribute values with parent values in the taxonomy. For continuous variables, microaggregation can be used, which is the aggregation (e.g., by applying the mean or median function) of the most similar observations.

When only a few records or attributes increase the reidentification risk significantly, it can be more efficient to suppress cells or whole records instead of using generalization. In general, one can decide to make a full domain or a local form of hiding details (recoding or generalization). In the former case, all records are affected, whereas in the latter case, a selection of records takes place, which reduces the distortion on the data, but increases the complexity of the anonymization process. Two methods that try to achieve a better balance between utility and risk by considering more than one attribute at a time are multidimensional generalization[31] and multidimensional suppression.[30]

Disassociating attributes

The central aim of methods in this group is to disassociate the relationship between QIDs and sensitive attributes. This is mainly done by scattering sensitive attributes in subrecords of the published data, which can guarantee the preservation of the original sensitive attribute values in the transformed data set, in contrast to generalization and suppression. However, such a guarantee also has a detrimental effect, because the original of the attribute values remain in the data, which increases the risk of membership disclosure. Names of the approaches for dissociating attributes are very similar, for example, *bucketization*,[32] *anatomization*,[21,33] *slicing*,[24] and *disassociation* in a narrow sense.[34] For example, bucketization partitions the tuples into buckets of similar sensitive attribute values and disassociates the sensitive attributes from the QIDs by randomly permuting the sensitive values in each bucket. All these methods pretend to have less utility loss than generalization and suppression and to handle high-dimensional data, for which the generalization and suppression would produce useless data if k-anonymity is required.

Perturbation of the data

Pertubation refers to the transformation of original data values.[10] Concretely, perturbation is the distortion of the data by adding noise (randomization), swapping values, aggregating values (e.g., using the mean of a group), or using synthetic data. This family of anonymization methods produces transformed data that have similar statistical information as the original data. The perturbed data records do not correspond to real-world entities, which makes attribute disclosure rather improbable, but makes subanalyses rather problematic because of the lack of truthfulness in the transformed data. Especially for noise addition, it is important to have a proper randomization scheme. This means, for example, that noise should have correlations; otherwise, the correlation between the attributes is not masked and can be used to infer some characteristics.[35]

Conclusion

The discussion of basic notions related to anonymization showed that there is some work to do to achieve a mutual understanding between legal and technical experts concerning some of these notions. With respect to sensitive attributes that are listed in texts of law, there is an agreement that they need special protection. However, it depends on further assumptions regarding the availability of external information on such sensitive attributes whether they are perturbed in the anonymization process or not. In addition, there are attributes that are deemed sensitive without being listed in a text of law. This might be based on good reasons but having criteria for classifying attributes as sensitive and making assumptions explicit seems desirable, although we are aware of the fact that the determination of such criteria can be very difficult, especially for context-dependent sensitivity.

Regarding identification, lacking mutual understanding can be ascribed to the problem of defining this term at all. We have tried to circumvent a circular definition by considering two decisions: first, deciding which identifier should stand for the individual and second, deciding how to recognize the individual (by someone she has met before). The first decision implies again some circularity, if the

association between the individual and the unique identifier is not based on natural identifiers such as the genome or the fingerprints. Such natural unique identifiers can be matched to the individual by taking samples from him or her. When using artificial unique identifiers, circularity seems to be unavoidable (how to be sure that one is only pretending to be the one to whom the identifier belongs?).

Besides notational questions, we also touched on some challenges with respect to anonymization techniques. The anonymization of genetic data aggravates these challenges because of the volume as well as the content of molecular data. Pakstis et al.[36] showed that a carefully chosen set of 45 single nucleotide polymorphisms is frequently sufficient to provide entity matches with a type 1 error of 10^{-15}. Erlich and Narayanan provide several examples for breaching genetic privacy.[37] From our perspective, the major problem of genetic data—besides the volume—stems from their inherent double nature, containing both sensitive attributes and unique identifiers (not only QIDs!). Enforcing k-anonymity for genetic data (e.g., sequences) will often lead to useless data, even if one uses state-of-the-art methods.[38] On the other hand, if an attacker can link an individual by using his or her genome, it would permit the adversary to learn many sensitive details, such as phenotypes (e.g., skin color) and susceptibility to diseases. Therefore, alternatives are necessary for providing a sufficient balance between utility and privacy for high-dimensional molecular data that are also relevant for other settings, for example, fully homomorphic encryption, secure multipart computation,[39–41] as well as tight access policies.

It is useful to have some hints regarding anonymization methods and techniques that are suitable for the specific purpose. For example, Templ et al.[42] suggest using generalization and local suppression (hiding some details in attributes) to achieve low disclosure risk when only a few QIDs are in the data set. In other more complex cases, methods that are more sophisticated should be applied. As the WP29 group states: ''anonymization should not be regarded as a one-off exercise, and the attending risks should be reassessed regularly by data controllers. The state of the art in methods and technology should be considered''.[43]

Finally, using a dialectical definition process to merge technical and legal perspectives on terms seems important for enhancing mutual understanding. For example, the term ''aggregated data'' can have the connotation of ''nonindividual,'' even if it allows identification of individuals. As this term is not used in the EU Data Protection Directive, it is crucial to define clearly—together with experts from different fields—what ''aggregated data'' mean. From a technical perspective, privacy-preserving data mining experts could make a very useful contribution to this, because they are addressing the problem of preventing identity disclosure on ''aggregated'' query results. A technical definition can then be the start point of a concerted definition process.

Author Disclosure Statement

No conflicting interests exist.

References

1. Research & Innovation. Science with and for Society—Policy. Available at: http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=science Accessed February 22, 2016.
2. grants.nih.gov. NIH Sharing Policies and Related Guidance on NIH-Funded Research Resources. Available at: http://grants.nih.gov/grants/sharing.htm Accessed February 22, 2016.
3. Wellcome Trust. Data sharing. . Available at: www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/ Accessed February 22, 2016.
4. Castellani J. Are clinical trial data shared sufficiently today? Yes. BMJ 2013;347:f1881.
5. Emam KE, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015;350:h1139.
6. TMF Comment on European Parliament Draft of General Data Protection Regulations. Available at: www.tmf-ev.de/Desktopmodules/Bring2Mind/DMX/Download.aspx?EntryId=25100&PortalId=0 Accessed February 22, 2016.
7. Willenborg L. *Elements of Statistical Disclosure Control*. New York, NY: Springer; 2013.
8. Hundepool A, et al. *Statistical Disclosure Control*. Chichester, UK: John Wiley & Sons; 2012.
9. El Emam K, Dankar FK. Protecting privacy using k-anonymity. J Am Med Inform Assoc 2008;15:627–637.
10. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey of recent developments. ACM Comput Surv 2010;42:14:1–14:53.
11. Malin BA, Emam KE, O'Keefe CM. Biomedical data privacy: Problems, perspectives, and recent advances. J Am Med Inform Assoc 2013;20:2–6.
12. Claerhout B, De Moor GJE. Privacy protection for HealthGrid applications. Methods Inf Med 2005;44:140–143.
13. Aggarwal CC, Yu PS. *Privacy-Preserving Data Mining: Models and Algorithms*. New York, NY: Springer US; 2008.
14. Westin AF. *Privacy and Freedom*. London, Sydney, Toronto: The Bodley Head Ltd.; 1970.
15. DIRECTIVE 95/46/EC. Available at: http://ec.europa.eu/justice/policies/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf Accessed February 22, 2016.
16. Emam KE, Álvarez C. A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques. Int Data Priv Law 2015;5:73–87.
17. Article 29 Working Party—European Commission. Available at: http://ec.europa.eu/justice/data-protection/article-29/index_en.htm Accessed February 22, 2016.
18. OECD Glossary of Statistical Terms—Quasi-identifier Definition. Available at: https://stats.oecd.org/glossary/detail.asp?ID=6961 Accessed February 22, 2016.
19. Ohno-Machado L, Vinterbo S, Dreiseitl S. Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance. J Am Med Inform Assoc 2002;9:S115–S119.
20. Summary of the HIPAA Privacy Rule. Available at: www.hhs.gov/ocr/privacy/hipaa/understanding/summary/ Accessed February 22, 2016.
21. Fung BCM. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Boca Raton: Chapman & Hall/CRC; 2010.
22. Gionis A, Tassa T. k-Anonymization with minimal loss of information. IEEE Trans Knowl Data Eng 2009;21:206–219.
23. Iyengar VS. Transforming data to satisfy privacy constraints. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2002: 279–288.

24. Li T, Li N, Zhang J, Molloy I. Slicing: A new approach for privacy preserving data publishing. IEEE Trans Knowl Data Eng 2012;24:561–574.
25. Dankar FK, Emam KE, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Med Inform Decis Mak 2012;12:66.
26. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data 2007;1(1): Article No. 3.
27. Article 29 WP Opinion 4/2007 on the concept of personal data. Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf Accessed February 22, 2016.
28. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data.* Baltimore: ACM; 2005: 49–60.
29. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. J Biomed Inform 2014;50:4–19.
30. Kisilevich S, Rokach L, Elovici Y, Shapira B. Efficient multidimensional suppression for k-anonymity. IEEE Trans Knowl Data Eng 2010;22:334–347.
31. LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional k-anonymity. In: *Proceedings of the 22nd International Conference on Data Engineering, 2006. ICDE'06*. Los Alamitos: IEEE; 2006: 25–25.
32. Cao J, Karras P, Kalnis P, Tan K-L. SABRE: A sensitive attribute bucketization and redistribution framework for t-closeness. VLDB J 2010;20:59–81.
33. Menzies T, Kocaguneli E, Turhan B, Minku L, Peters F. *Sharing Data and Models in Software Engineering*. Waltham, MA: Morgan Kaufmann; 2014.
34. Loukides G, Liagouris J, Gkoulalas-Divanis A, Terrovitis M. Disassociation for electronic health record privacy. J Biomed Inform 2014;50:46–61.
35. Huang Z, Du W, Chen B. Deriving private information from randomized data. In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data.* New York: ACM; 2005: 37–48.
36. Pakstis AJ, Speed WC, Kidd JR, Kidd KK. Candidate SNPs for a universal individual identification panel. Hum Genet 2007;121:305–317.
37. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet 2014;15:409–421.
38. Li G, Wang Y, Su X. Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices. Comput Methods Programs Biomed 2012;108:1–9.
39. He D, et al. Identifying genetic relatives without compromising privacy. Genome Res 2014;24:664–672.
40. Ayday E, Raisaro JL, McLaren PJ, Fellay J, Hubaux J-P. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. In: *Proceedings of the 2013 USENIX Conference on Safety, Security, Privacy and Interoperability of Health Information Technologies*. Berkeley, CA: USENIX Association; 2013:1–10.
41. Ayday E, Cristofaro ED, Hubaux J-P, Tsudik G. The chills and thrills of whole genome sequencing. Computer 2013; 99:1.
42. Templ M, Kowarik A, Meindl B. sdcMicro: Statistical Disclosure Control methods for anonymization of microdata and risk estimation. Available at: http://cran.r-project.org/web/packages/sdcMicro/index.html Accessed February 22, 2016.
43. Opinions and Recommendations—Justice. Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/index_en.htm Accessed February 22, 2016.

Address correspondence to:
Murat Sariyar, Dr rer physiol
TMF (Technologie- und Methodenplattform e.V.)
Berlin 10117
Germany

E-mail: murat.sariyar@charite.de